**GSS Data**

Chris Sinclair

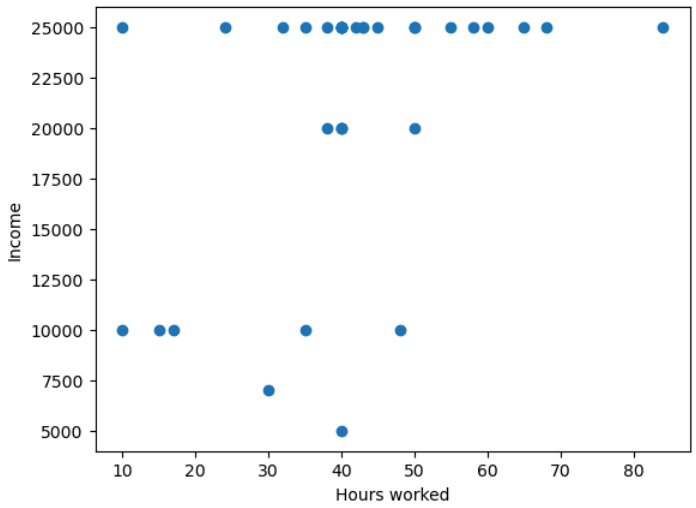
February 23rd, 2023

*(week6gss.py)*

For the GSS data, I chose the year of 1989 as it's my birth year. The first task encompassed plotting out the Hours Worked and Income data points to see if there was any correlation. I had to do quite a bit of cleaning on this, starting with getting rid of essentially null values in the Hours Worked column by removing items with certain strings, such as "Inapplicable" or "No answer". Then I had to convert the data type to float so it would be easier to calculate.

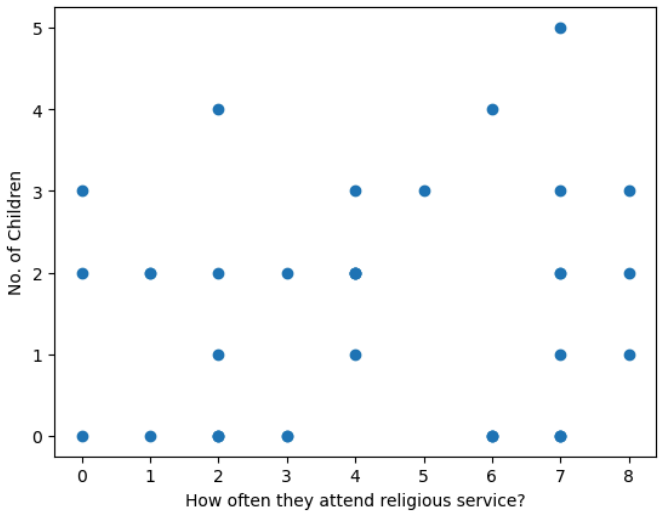
Next, the Income column had similar null values as well ("Do not Know" and "Refused") that I got rid of. It also specified certain ranges of income, with the entries outlined as: [$5000-$5999], [$7000- $7999], [$10,000-$14,999], [$20,000-$24,999] and [$25,000 or more]. I thought this was particularly weird because these ranges ignored the numbers $6000-$6,999 as well as $8,000-$9,999. Either way, I converted each of these to a float value by only taking the first number and assuming the minimum of the range for the entry. Thus, an entry of $5,000-$5,999 was now seen by the program simply as 5000.

Finally, I plotted these data points against each other but had trouble cleaning it up to specify the initial ranges in the final graph for the y-ticks. Regardless, the plot is pictured below:




I also ran an analysis to find the correlation coefficient was 0.43. R-squared was only 0.18. Looking at the plot above as well as these numbers, it's pretty clear that there is a pretty weak correlation between these data points. Some extra cleaning up I could do to tighten those numbers up would be to get rid of certain outliers, such as the one that shows Hours Worked above 80, as well as the other one on the opposite end that shows Hours Worked at 10 but with an Income above $25,000. I still don't think getting rid of those would greatly affect the correlation though, so I don't think there's much of a correlation at all.

Task #2 was similar. I created a subset table and converted each of two relevant columns (childs and attend), to floats. I didn't really see any outliers in this subset, nor were there any null values, so there wasn't much cleaning necessary at this point. I plotted the data points and, at least visually, didn't see much of a correlation.




However, I still had to run a correlation analysis. The correlation coefficient between these data points was only 0.097, which shows a very weak correlation. Furthermore, R-squared was an incredibly low 0.009, which further confirms how unrelated these data points are. It seems as though there isn't really any data that determines if larger families attend church services more or less. Putting this question up against a different kind of data point though, such as religious upbringing, would probably show a much higher correlation.

## **Real Estate Data**

*(realestate.py)*

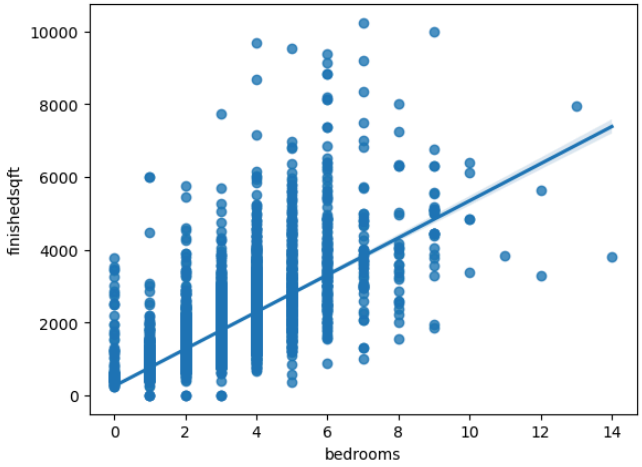
For the Real Estate data, I first had to create a subset table consisting of only the relevant columns as well as eliminating outliers. With the "Rooms Built" column, I decided to cap it at 15 rooms built per house each year since I noticed that anything above that was bringing up the total rows count very slightly.

Task #1 of the Real Estate section involved conducting a correlational analysis of the number of rooms in a house and the year the house was built. The correlation coefficient was -0.37, which shows a relatively weak correlation between those two fields. R-squared came out to 0.14, which confirms that there is an incredibly weak correlation between these two data points.

As for Task #2, I got rid of the outlier houses that contained more than ten (10) bathrooms from my subset table and ran a linear regression analysis to get a correlation coefficient of 0.55,

which shows a moderate positive correlation between the Last Sold Price and the Total Number of Bathrooms. R-squared came out to 0.3 though, which means that, since it's closer to 0 than to 1, the model isn't a strong foundation to explain the data. We would probably need to look at other variables to present a stronger case of correlation.

For Task #3, I again had to clean it up a bit by getting rid of outliers for Total Bedrooms and Finished Square Feet. Thankfully, in all of these Real Estate data analyses, there were no nulls to clean up or fields that needed altered data types. Either way, the plot with a regression line for Task #3 is pictured below:




My interpretation is that, with more square footage comes more bathrooms. Out of curiosity, I decided to run a correlation coefficient to find that it comes out to 0.73 – the strongest correlation we've seen yet! However, since it's not between the range of 0.8 - 1.0, it is still considered only a moderate positive correlation. R-squared also comes out to a higher number than the others we've seen – 0.53. This number could definitely be higher to verify the correlation, but of all the data points we've explored so far, this one is still the highest by a significant margin.

## Final Thoughts

This milestone was definitely more of a challenge as I feel like my scatter plots, or at least configuring them in a specific way, is the toughest part right now. I made some progress here and there after converting items to floats. I'm not positive, but it also seems like sorting a column in the subset table also helps when trying to visualize it in a plot. The main area that really sticks it to me on this is the income range for Task #1 of the GSS data. I converted those ranges to the minimum floats with the intention of at least labeling the y-ticks as well as setting y-tick ranges in the plot to essentially work around the issue, but I had a hard time figuring that out.